

# Weekly Report

Pingping Shang

2013.11.11~2013.11.17

## 本周工作

- 1、将 C# 版本的熵聚类算法重构为 java 版本, 然后改进该算法用于年龄属性的分段。该熵聚类算法用于类别型数据或数值型数据。类别型数据会被转化为 int 型, 如 “Red” 记为 0, “Blue” 记为 1, 依次类推, 然后根据出现频率计算熵值。

样例(对 4 tuples 聚类):

	Color	Size	Texture		
[0]	Red	Medium	Hard	→	[0] 0 1 0
[1]	Red	Small	Soft		[1] 0 0 1
[2]	Blue	Large	Hard		[2] 1 2 0
[3]	Blue	Medium	Hard		[3] 1 1 0

对应到年龄分段任务 (对 age 聚类):

	商品 1	商品 2	商品 3	...
[age1]	买	否	买	...
[age2]	否	否	买	...
[age3]	买	买	否	...

(对算法的改进地方就是我们不是自由的聚类, 而是类似于分段)

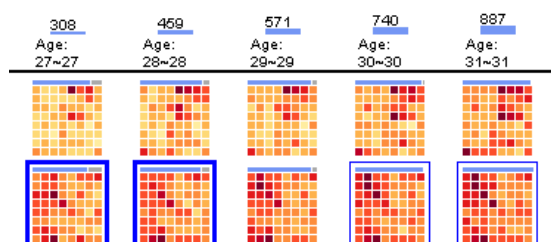
- 2、1) 实现 distance metric learning 算法, 我们只考虑了下面公式中  $A$  为对角阵的情况:

$$d(x, y) = d_A(x, y) = \|x - y\|_A = \sqrt{(x - y)^T A (x - y)}.$$

$$\begin{aligned} \min_A \quad & \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2 \\ \text{s.t.} \quad & \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A \geq 1, \\ & A \succeq 0. \end{aligned}$$

该问题整理后可认为是有线性约束的非线性规划问题。实现过程中先用 matlab 描述并解答问题, 再 build 为 jar 包, 在 java 程序中调用。

- 2) 该问题用在标签项目中的意义为: 用户选择相近的 block, 两个 block 为一组, 可选择多组 (见下图), 然后这些相近信息作为约束条件输入, 目标为找到  $A$  (可认为商品的权重) 使得这些 block 组尽可能相近。



## 下周工作

考虑距离学习得到的距离定义如何使用，熵聚类如何用在标签的顺序选择。